Homework 1 **CSCE 421**

Due: 5.00 pm on September 19, 2019

Instructions for homework submission

a) For the **math problems**, please typewrite your answers in Latex, or handwrite your solution *very clearly*. Non-visible solutions will not be graded: we wouldn't like our TA to have to guess what you are writing :)

b) For the **experimental problems**, please write a brief report. At the end of the report, please include your code. Print the report, including the code.

c) **Staple all your answers and hand them out in paper in class or during office hours.**

d) Please start early :)

e) The maximum grade for this homework, excluding bonus questions, is **9 points** (out of 100 total for the class).

Question 1 (3 points)

Machine learning definitions: Our purpose is to create a coin classification system for a vending machine. What types of Machine Learning, if any, best describe the following scenarios? Please provide a brief explanation.

(1 point) (i) The exact specifications of each coin are measured by an engineer. The vending machine recognizes a given coin based on these specifications.

(1 point) (ii) An algorithm is presented with a large set of labeled coins and uses this data to infer decision boundaries, based on which the vending machine classifies new coins.

(1 point) (iii) An algorithm is successively presented with coins. Each time the algorithm makes a decision about the coin type and checks the correctness of the decision with the engineer. Based on the engineer's answer, the algorithm refines the process with which it makes the decision for the next coin.

Question 2 (7 points)

Classifying benign vs malignant tumors: We would like to classify if a tumor is benign or malign based on its attributes. We use data from the following UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)`. Inside “Homework 1” folder on Piazza you can find three files including the train and test data (named “hw1_ question1_train.csv”, “hw1_ question1_dev.csv”, and ”hw1_ question1_test.csv”) for our experiments. The rows of these files refer to the data samples, while the columns denote the features (columns 1-9) and the class variable (column 10), as described bellow:

1. Clump Thickness: discrete values $\{1, 10\}$
2. Uniformity of Cell Size: discrete values $\{1, 10\}$
3. Uniformity of Cell Shape: discrete values $\{1, 10\}$
4. Marginal Adhesion: discrete values $\{1, 10\}$
5. Single Epithelial Cell Size: discrete values $\{1, 10\}$

6. Bare Nuclei: discrete values $\{1, 10\}$
7. Bland Chromatin: discrete values $\{1, 10\}$
8. Normal Nucleoli: discrete values $\{1, 10\}$
9. Mitoses: discrete values $\{1, 10\}$
10. Class: 2 for benign, 4 for malignant (this is the **outcome**)

(a.i) (1 point) Data exploration: Using the training data, compute the number of samples belonging to the benign and the number of samples belonging to the malignant case. What do you observe? Are the two classes equally distributed in the data?

(a.ii) (1 point) Data exploration: Using the training data, plot the histogram of each feature (i.e., 9 total histograms). How are the features distributed in the 1-10 range? Are the sample values distributed equally for each feature?

(a.iii) (1 point) Data exploration: Randomly select 5 pairs of features. Using the training data, plot scatter plots of the selected pairs (i.e., 5 total scatter plots). Use a color-coding to indicate the class in which the samples belong to (e.g., blue for benign, red for malignant). What do you observe? How separable do the data look?

(b.i) (2 points) Implement a K-Nearest Neighbor classifier (K-NN) using the euclidean distance ($l2$-norm) as a distance measure to classify between the benign and malignant classes. **Please implement K-NN and do not use available libraries.**

(b.ii) (1 point) Explore different values of $K = 1, 3, 5, 7, \dots, 19$. You will train one model for each of the ten values of K using the train data and compute the classification accuracy (Acc) and balanced classification accuracy ($BAcc$) of the model on the development set. Plot the two metrics against the different values of K. Please report the best hyper-parameter K_1 based on the Acc metric, and the best hyper-parameter K_2 based on the $BAcc$ metric. What do you observe? **Please implement this procedure from scratch and do not use available libraries.**

Hint: $Acc = \frac{\textit{\# correctly classified samples}}{\textit{\# samples}}$

$BAcc = \frac{1}{2}\left(\frac{\textit{\# correctly classified samples from Class 1}}{\textit{\# samples in Class 1}} + \frac{\textit{\# correctly classified samples from Class 2}}{\textit{\# samples in Class 2}}\right)$

(b.iii) (1 point) Report the Acc and $BAcc$ metrics on the test set using K_1 and K_2. What do you observe?

(b.iv) (Bonus, 2 points) Instead of using the euclidean distance for all features, experiment with different types of distances or distance combinations, e.g. $l0$-norm or cosine similarity. Report your findings.